

APPLICATION
FOR
UNITED STATES DESIGN PATENT

TITLE: REMOTE DATA FACILITY OVER AN IP NETWORK
APPLICANT: ALEX VEPRINSKY, DAVID MEIRI AND RAMPRASAD SHETTY

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL445349548US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the U.S. Patent and Trademark Office, P.O. Box 2327, Arlington, VA 22202.

November 28, 2001
Date of Deposit

Signature

Leroy Jenkins
Typed or Printed Name of Person Signing Certificate

REMOTE DATA FACILITY OVER AN IP NETWORK

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority from U.S. Provisional Patent Application Ser. No. 60/325,658, filed September 27, 2001, incorporated herein by reference in its entirety for all purposes.

15
16
17
18
19
20
21
22
23
24
25

BACKGROUND

The invention relates generally to data storage systems, and in particular, to data storage systems with remote data mirroring capability.

Given the importance of the availability of information, several techniques have been developed for providing enhanced reliability and availability of data stored in a data storage system. One such technique is remote data mirroring. In a "mirrored" system, the data stored on one data storage system is replicated on another data storage system, preferably at a geographically remote site. Thus, if one or more storage devices on one of the data storage systems fails, or a catastrophic system failure should occur, the data is readily available in the form of a mirrored copy from the remote data storage system.

Devices on a data storage system (or source) that are mirrored on the same remote, data storage system (or target) are referred to as a device group. Likewise, devices on the target that serve or mirror devices on the same source are referred to as a device group. Device groups are used, for example, to insure the consistency of blocks of data too large to be stored in a single area, during planned or unplanned interruptions such as link failures or planned back-ups. Device groups typically span a number of physical and logical volumes, and, in some instances, as when data striping is used, several device groups may be mapped to different portions of a physical or logical volume.

Typically, in a remote data mirroring environment, the source and target device groups are arranged as pairs on any two systems and the source/target device group pairs are connected by dedicated data links (e.g., ESCON links) or switched data links (e.g., switched Fibre Channel links). The data links support writes during data backup operations and reads during data recovery operations.

Such point-to-point and switched data link topologies have distance limitations. To negotiate long distances, the ESCON (or Fibre Channel) data links (connected to data ports of

5 the data storage systems, local and remote) are coupled via a leased line (such as T3) or an IP network. There are significant drawbacks associated with these types of long distance solutions, however. For example, the T3 link is extremely expensive and very slow compared to the ESCON or Fibre Channel links. In addition, because connections using these solutions span diverse network protocols and interfaces, some type of adapter box must be used to translate
10 between diverse protocols and interfaces of, say, ESCON and T3, or ESCON and IP. Typically, the adapter box is designed, configured and managed by some entity other than the data storage system supplier. This means that some aspects of the data storage system's performance are either dictated by the adapter box (for example, delays due to the buffer constraints or encapsulation, availability of IP services), or the quality of the IP line, for example, an IP
15 connection provided by an Internet Service Provider, and are therefore not within the control of the data storage system supplier. Moreover, the design of the adapter boxes, in particular, those supporting FC-to-IP services, can be quite complex, making networks of remotely-connected data storage systems that employ such boxes expensive from a field service perspective.

SUMMARY

In one aspect, the invention provides methods and apparatus, including computer program products, for operating a data storage system in a remote data mirroring arrangement of data storage systems. The methods include determining that storage traffic is to be transferred between the data storage system and a remote data storage system to which the data storage
25 system is coupled by an IP network and enabling transfer of the storage traffic between the data storage system and the remote data storage system over the IP network using a native connection to the IP network.

Particular implementations of the invention may provide one or more of the following advantages.

The present invention allows data storage systems in a remote, data mirrored configuration to participate directly as members of and use the full set of services of an IP network. Allowing the data storage systems to establish native connections to an IP network directly eliminates the need for expensive third-party adapter boxes, which limit the extent to which the data storage systems can use the services of an IP network. Moreover, because the
35 adapter boxes are eliminated, a data storage system supplier is able to better control and monitor

5 performance of remote data service functions that use an IP network (such as the Internet) for long distance transfer of storage traffic.

Other features and advantages of the invention will be apparent from the following detailed description and from the claims.

5

DESCRIPTION OF DRAWINGS

FIG. 1 is block diagram of a data processing system including host computers coupled to a data storage system, which includes storage devices coupled to a storage controller for controlling data transfers between the host computers and storage devices as well as between the data storage system and another, remote data storage system.

10

FIG. 2 is a detailed block diagram of the data storage system and its storage controller (shown in FIG. 1), which includes a remote (Remote Data Facility or “RDF”) director for managing the exchange of RDF storage traffic between the data storage system and a remote data storage system over an IP network.

 15
 20
 25

FIG. 3 is a block diagram of a remote, data mirrored arrangement of data storage systems (like the one depicted in FIGS. 1 and 2) that are interconnected by an IP network (shown as the Internet) and are capable of sending storage traffic to each other over the IP network using native connections.

FIG. 4 is a block diagram of a two-processor implementation of the remote director (of FIG. 2) to enable native connections to an IP network.

FIG. 5 is a depiction of the software executed by the processors in the remote director.

FIG. 6 is another block diagram of the remote director that shows details of a shared memory implementation for exchanging socket interface messages across processor boundaries.

Like reference numerals will be used to represent like elements.

25

DETAILED DESCRIPTION

Referring to FIG. 1, a data processing system 10 includes host computers 12a, 12b,..., 12m, connected to a data storage system 14. The data storage system 14 can be, for example, that made by EMC Corporation and known as the Symmetrix data storage system. The data storage system 14 receives data and commands from, and delivers data and responses to, the host computers 12. The data storage system 14 is a mass storage system having a controller 16 coupled to pluralities of physical storage devices shown as disk devices 18a, disk devices 18b, ..., disk devices 18k. Each of the disk devices 18 is logically divided, in accordance with known techniques, into one or more logical volumes.

30

The controller 16 interconnects the host computers 12 and the disk devices 18. The controller 16 thus receives write commands from the various host computers over buses 20a,

35

5 20b, ..., 20m, respectively, for example, connected and operated in accordance with a SCSI
protocol, and delivers the data associated with those commands to the appropriate devices 18a,
18b, ..., 18k, over respective connecting buses 22a, 22b, ..., 22k. Buses 22 also operate in
accordance with a SCSI protocol. Other protocols, for example, Fibre Channel, could also be
used for buses 20, 22. The controller 16 also receives read requests from the host computers 12
10 over buses 20, and delivers requested data to the host computers 12, either from a cache memory
of the controller 16 or, if the data is not available in cache memory, from the disk devices 18.

In a typical configuration, the controller 16 also connects to a console PC 24 through a
connecting bus 26. The console PC 24 is used for maintenance and access to the controller 16
and can be employed to set parameters of the controller 16 as is well known in the art.

15 The controller may be connected to a remote data processing system like the data
processing system 10 or a remote data storage system like the data storage system 14 (shown in
dashed lines) for data back-up capability by a data link 28. The data link 28 is implemented
according to Gigabit Ethernet protocols. Other network protocols can be used as well. The data
link 28 enables a remote data storage system to store on its own devices a copy of information
stored in the devices 18 of the data storage system 14 in a mirrored manner, as will be described.

20 In operation, the host computers 12a, 12b, ..., 12m, send, as required by the applications
they are running, commands to the data storage system 14 requesting data stored in the logical
volumes or providing data to be written to the logical volumes. Referring to FIG. 2, and using
the controller in the Symmetrix data storage system as an illustrative example, details of the
25 internal architecture of the data storage system 14 are shown. The communications from the host
computer 12 typically connect the host computer 12 to a port of one or more host directors 30
over the SCSI bus lines 20. Each host director, in turn, connects over one or more system buses
32 or 34 to a global memory 36. The global memory 36 is preferably a large memory through
which the host director 30 can communicate with the disk devices 18. The global memory
30 includes a common area 38 for supporting communications between the host computers 12 and
the disk devices 18, a cache memory 40 for storing data and control data structures, and tables 42
for mapping areas of the disk devices 18 to areas in the cache memory 40.

35 Also connected to the global memory 36 are back-end (or disk) directors 44, which
control the disk devices 18. In the preferred embodiment, the disk directors are installed in the
controller 16 in pairs. For simplification, only two disk directors, indicated as disk directors 44a

5 and 44b, are shown. However, it will be understood that additional disk directors may be employed by the system.

10 Each of the disk directors 44a, 44b supports four bus ports. The disk director 44a connects to two primary buses 22a and 22b, as well as two secondary buses 22a' and 22b'. The buses are implemented as 16-bit wide SCSI buses. As indicated earlier, other bus protocols besides the SCSI protocol may be used. The two secondary buses 22a' and 22b' are added for redundancy. Connected to the primary buses 22a, 22b, are the plurality of disk devices (e.g., disk drive units) 18a and 18b, respectively. The disk director 44b connects to two primary buses 22c and 22d. Connected to the primary buses 22c, 22d are the plurality of disk devices or disk drive units 18c and 18d. Also connected to the primary buses 22c and 22d are the secondary buses 22a' and 22b'. When the primary bus is active, its corresponding secondary bus is inactive, and vice versa. The secondary buses of the disk director 44b have been omitted from the figure for purposes of clarity.

15 20 Like the host directors 20, the disk directors 44 are also connected to the global memory 36 via one of the system buses 32, 34. During a write operation, the disk directors 44 read data stored in the global memory 36 by a host director 30 and write that data to the logical volumes for which they are responsible. During a read operation and in response to a read command, the disk directors 44 read data from a logical volume and write that data to global memory for later delivery by the host director to the requesting host computer 12.

25 As earlier mentioned, the data storage system 14 can be remotely coupled to another data storage system 14 in a mirrored storage configuration, using the data link 28. Still referring to FIG. 2, each data storage system 14 in the mirrored storage configuration includes a remote director 48 to connect to the data link 28 and handle transfers of data over that link. The remote director 48 communicates with the global memory 36 over one of the system buses 32, 34.

30 Referring to FIG. 3, a remote data services (e.g., data mirroring) storage configuration 50 includes two or more of the data storage systems 14 (illustrated as three data storage systems 14a, 14b and 14c). The data storage systems 14a, 14b and 14c are directly coupled to an IP network (shown as the Internet 52) by respective data links 28a, 28b and 28c. The data links 28 are implemented as Gigabit Ethernet transmission channels as mentioned earlier, but any suitable transmission medium for supporting TCP/IP traffic may be used. The data links 28, and the IP network 52, are used to support connections for carrying TCP/IP traffic between the units 14.

5 For example, a first connection 54a may be established between the data storage systems 14a and 14b. A second connection 54b may be established between the data storage systems 14b and 14c. A third connection 54c may be established between the data storage systems 14c and 14a. In the system 50, the data storage systems 14 are configured for remote data mirroring capability. More specifically, in the example shown, there are eight device groups, S1, S2, S3, S4, T1, T2, 10 T3, T4, which are indicated by reference numerals 56a, 56b, 56c, 56d, 56e, 56f, 56g, 56h, respectively. Four of the device groups, S1 through S4, are source device groups, and device groups T1 through T4 are target device groups. In the example shown, the data storage systems 14 are configured in the following manner: the data storage system 14a supports device groups S1, S2 and T3; the data storage system 14b supports device groups S4, T1 and T2; and the data storage system 14c supports the device groups S3 and T4. Thus, the devices in the source group 15 S1 are mirrored in the devices in corresponding target device group T1, devices in the source group S2 are mirrored in the devices in corresponding target device group T2, and so forth. Thus, the units use TCP/IP to exchange storage traffic as required by remote data facility services, for example, the data storage systems 14a and 14b establish a connection with each other so that the data storage system 14a can provide a copy of data residing on the source device group S1 to the target device group T1. Thus, the architecture of the remote directors 48 (as will be described) in the each of the data storage systems 14 allows those systems to use the Internet infrastructure for disaster recovery and other remote data services. Although the IP network 52 is shown as the public Internet, it could instead be a private network.

20 As shown in FIG. 4, the remote director 48 includes an RDF director 60 and a link director 62. The RDF director 60 includes a processor 64 coupled to a local, nonvolatile memory (NVM) 66. The NVM 66 includes a control store 68 and a parameter store 70. The link director 62 includes a processor 72 coupled to its own, NVM 74, which also includes a control store 76 and a parameter store 78. The directors 60, 62 each have access to a shared memory 80. The processor 64 controls the overall operations of the RDF director 62 and communications 25 with the memories 66 and 80. The control store 68 stores firmware (or microcode) 82 and parameter store stores parameter data, both of which are read each time the data storage system 14 is initialized. The microcode 82 is copied into the control store 68 at initialization for subsequent execution by the processor 64. The processor 72 controls the overall operations of 30 the link director 62 and communications with the memories 74 and 80. The control store 76

5 stores link firmware (or microcode) 84 and the parameter store 78 stores parameter data, both of which are read each time the data storage system 14 is initialized. The microcode 84 is copied into the control store 76 at initialization for subsequent execution by the processor 72.

Referring to FIG. 5, the microcodes 82 and 84 are shown. The RDF director's microcode 82 includes an RDF emulation layer 94, a Common Device Interface 96 and a first socket relay layer 98. The microcode 84, executed by the link processor 72, includes a second socket relay layer 100, a TCP/IP layer 102 and a network driver 104. Collectively, the socket relays 98, 100 represent a socket interface 108, and pass socket messages to each other. Although the interface 108 between the higher-level RDF emulation/CDI layers (which execute on the emulation processor 64) and the TCP/IP protocols of layer 102 (which execute on the link processor 74) is shown as being implemented as a socket interface, other interfaces could be used for communications between the RDF emulation and the TCP/IP protocols software.

The RDF emulation 94 can include the following: a system calls layer 110; advanced functionality modules 112, which may be optional at the director level or even at the data storage system level; common function modules 114, which are provided to each director in the controller 16; and an interface (director application) module. Interface modules exist for each of the different types of directors that are available based on connectivity and/or function, for example, a Remote Data Facility (RDF) interface defines the functionality of the remote director 48, mainframe and Open Systems host interfaces, respectively, define host directors 30, and a back-end interface defines the functionality of the back-end director 44.

25 The emulation is defined as software that implements both an Upper Level Protocol (ULP), that is, a protocol associated with functionality in one or more of layers 110, 112 and 114 (from FIG. 5), and functions corresponding to the RDF interface 116. Thus, the emulation 94 resides above any physical transport layers and includes software corresponding to the RDF interface 114 as well as software implementing a ULP.

30 The CDI 96 recognizes that different physical transports have different physical formats, data capacities and access characteristics. Consequently, the CDI 96 accommodates and isolates those physical transport differences so that those portions of the drivers and emulations that interact with each other are generic in nature. The CDI 96 provides for versatility and is intended to support any existing or envisioned transport functionality (or protocol). In addition to abstracting the details of different physical transport protocols, the CDI handles physical data

5 movement (e.g., via a DMA mechanism, as described below) and makes that data movement
transparent to emulation software.

The CDI can be viewed as being embodied in an I/O control block (hereinafter, "IOCB")
data structure. This IOCB data structure is a generic structure that serves to define a common
interface between the emulation 94 and a CDI compliant lower layer (CDI driver) with which the
10 emulation 94 communicates in transferring commands and data. To make a request (containing a
ULP command) to a CDI driver, the RDF emulation 94 uses a call, 'CDI IOCTL' that takes as its
only parameter a pointer to an IOCB describing the request. During the lifetime of that request
and its associated IOCB, the control of the IOCB alternates between the emulation and the CDI
driver that has accepted it. The CDI driver has control of the IOCB while an IOCTL call is
15 outstanding. The RDF emulation 94 has control of the IOCB when the call request has been
completed. Notification of events, e.g., the completion of an IOCTL call or the arrival of a new
ULP command, is signaled by the CDI driver to the emulation by placing corresponding IOCBs
on queues referred to herein as event (or completion) queues. Thus, the emulation detects a call
request completion status when it determines that the IOCB associated with the call has been
placed on an event queue by the CDI driver. By removing the IOCB from the event queue, the
emulation gains control of the buffer that had been allocated to that IOCB.

The CDI 96 may be supported in a polled or interrupt driven environment. In a polled
environment, the emulation must make periodic calls to a routine that acts as an interrupt service
routine in that is gives the driver a chance to look at the physical interface and process any
20 accumulated events. This call must be made frequently to facilitate the timely discovery of new
events or the completion of requests. In an interrupt driven environment, interrupts allows
events to be processed as they occur.

Further architectural and implementation-specific details of the CDI 96 can be found in
co-pending U.S. patent application Ser. No. 09/797,347, filed March 1, 2001, incorporated herein
30 by reference.

Still referring to FIG. 5, below the CDI 96 is the socket interface 100. In the described
embodiment, the RDF emulation 94 and the socket interface 100 have knowledge of the CDI
format. Thus, the CDI 96 serves to isolate the RDF emulation 94 from the TCP/IP layer.

Implementation-specific details of the TCP/IP layer 102, as well as lower network layers
35 104, 106 are implemented in known fashion and therefore described no further herein. It will be

5 appreciated that one skilled in the art would be able to implement the required link processor software (as well as any special hardware assists, e.g., DMA, not shown) necessary to transfer and receive packets over a Gigabit Ethernet data link using TCP/IP.

10 Although FIG. 5 shows the link processor firmware 84 as including network (e.g., Gigabit Ethernet) driver and hardware interface software (layers 104, 106), it will be appreciated that one or both of these layers could be implemented in a separate, commercially available Gigabit MAC device or chipset.

15 Referring to FIG. 6, a conceptual depiction of the interface 48 that shows some details of the shared memory 80 used for passing socket messages between the emulation processor 64 and the link processor 72 is shown. The shared memory 80 includes data structures for messages 120 and data 122, respectively. The messages are message related to establishing and tearing down individual TCP/IP connections. The data is the data to be encapsulated in a TCP/IP protocol data unit and passed down the protocol stack for processing and transmission over the Gigabit Ethernet data link, or data that was received over the link and decapsulated/processed as it is passed up the protocol stack in known fashion. The message data structures include outgoing and inbound data structures, 120a and 120b, for outgoing and inbound messages, respectively. Likewise, the data structures for managing transfer of data also include an outgoing data structure 122a and an inbound data structure 122b. All of the structures 120a, 120b 122a, 122b may be implemented as the same type of data structure, for example, circular rings.

20 It will be appreciated that the director 48 has been implemented as a two-processor architecture for performance reasons, that is, to off load the processing intensive TCP/IP operations from the processor that handles the RDF interface to the link processor. However, a single processor solution is also contemplated.

25 In addition, while the embodiment described above passes socket messages across the two-processor boundary, it may be possible to split the CDI between processors so that the messages that are passed between processors are CDI messages instead of socket messages. Such an implementation would require that the TCP/IP layer have knowledge of and be coded to conform to the CDI.

30 It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the

5 scope of the invention, which is defined by the scope of the appended claims. Other embodiments are within the scope of the following claims.

What is claimed is:

11